Human Disease Network Guided Discovery of Interesting Itemsets in Hospital Discharge Data

Gregor Stiglic
Faculty of Health Sciences, University of Maribor
Zitna ulica 15
2000 Maribor, Slovenia
gregor.stiglic@uni-mb.si

ABSTRACT
Standard knowledge discovery techniques, such as unsupervised or supervised descriptive rule discovery, have been widely used in medical data mining. Most of the research is focused on developing effective association rule evaluation metrics that would allow discovery of exceptional and interesting patterns. This study tries to integrate information on comorbidity obtained from recently very popular human disease networks, to rule learning from large medical datasets. The proposed approach is presented in a novel application of age related itemset mining from hospital discharge data. Such approach allows discovery of emerging patterns based on the age of patients that can be used to identify the age groups with the increased risk of comorbidities.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms
Algorithms

Keywords
human disease networks, itemset discovery, association rules

1. INTRODUCTION
In the recent years we have witnessed an increased amount of studies focusing on complexity of relations and co-occurrence of multiple diseases. Datasets containing large numbers of patients allow us to observe high number of co-occurring diseases on thousands of patients. A recent study by Steinhaeuser and Chawla [6] points out that our health care system is mostly reactive, meaning that we have become proficient at diagnosing diseases and developing treatments to cure them or prolong the life of a patient. However, we should now put more focus on proactive care aiming especially at prediction of the disease-related risks a few years before they actually happen and guide the patient to avoid them, instead of just curing them. A lot of information on how to achieve this is hidden in large amounts of available clinical data that should be used to discover new patterns that have not been discovered by classical statistical techniques or early data mining approaches.

One of the first approaches using large number of patients to construct disease related networks was a study by Hidalgo et al. [4] where authors demonstrated the usefulness of human disease networks to study the properties of co-occurring diseases. The study also demonstrates the potential of phenotypic data in form of the human disease network to complement genotypic and proteomic datasets that were extensively studied and analyzed in the past. The main contribution of the mentioned study that was also used in our research is the introduction of two metrics for quantification of comorbidity relationships.

An approach that exploits the information obtained by constructing an interconnected network of human diseases to speed up the interesting itemsets discovery is presented in this paper. Additionally, we introduce a novel approach to detection of interesting age group related itemsets from hospital discharge data. The following section presents the basics of interesting itemset discovery and is followed by Section 3 that introduces the methodology used in our experiments. Section 4 describes the results of the experiments and is followed by the last section with conclusions and future work.

2. INTERESTING ITEMSET DISCOVERY
Itemsets, representing co-occuring items in data, have been historically used in the association rules terminology since their introduction by Agrawal, Imielinski and Swami [1]. In some cases itemsets can be used to describe interesting patterns by themselves, without the typical left hand side and right hand side of the rule, typically found in association rules. In most cases, where itemset discovery should be used, the user would be misled into believing that there is a causal relationship between the left and the right hand side of the association rule. As already stated in previous work by Webb [8], there is a large amount of literature on association rules [1, 3, 5], but a lot less on interesting itemset discovery. Usually itemset discovery aims to find itemsets that co-occur more frequently than expected [8].

We will define a problem of interesting itemset discovery applied to hospital discharge data. Therefore we use the notation that introduces diagnoses instead of items and patients instead of transactions, which are usually met in definitions of association rule discovery. Let $I = \{d_1, d_2, \ldots, d_n\}$ be a set of n binary attributes representing presence of medical diagnoses. Let $D = \{p_1, p_2, \ldots, p_m\}$ be a set of patient records called the database. Each patient in D contains a patient id (PID) and a subset of diagnoses I.

KDD-DMH'11, August 21, 2011, San Diego, California, USA.

The cover of an itemset I, $cov(I)$ is the set of $PIDs$ for the patient records that contain I:

$$cov(I) = \{i : 1 \leq i \leq m \wedge I \subseteq p_i\}$$

The support of an itemset I is defined as the proportion of patient records in dataset D covered by I:

$$\sup(I) = \frac{|cov(I)|}{n}$$

Usually support of an itemset is used to evaluate the interestingness of and itemset, where we aim to discover itemsets that demonstrate unusual support patterns. In our study an application to hospital discharge data is used to demonstrate the novel measure of itemset interestingness that can be used for age related itemset discovery. Additionally to PID and a subset of diagnoses I, each patient record in our database contains information on the age of a patient at the time of discharge from the hospital. This allows age related evaluation of rules. More precisely, we aim to find itemsets exhibiting higher than expected support for a given age group. Width of age window that is used to define age group can be set by the user of the proposed interesting itemset discovery method.

Evaluation of itemset interestingness is done by comparing the support of I in the age group A_i with the support for I outside the A_i for all age groups A_i; $a_{min} \leq i \leq a_{max}$. We basically compare proportions of two independent groups; therefore a two-tailed z-test was used to compare supports. Maximal z value is used to define the interestingness of each itemset. Additional to the z-value it is also important to observe the age group intervals of the most interesting solutions that define the age groups where the risk of the target disease significantly increases when specific diagnoses co-occur.

2.1 Statistical Significance

Due to high number of discovered itemsets during the search process, we should adjust the significance level of z-test accordingly. In our approach we propose the adjustment of critical values to reduce the risk of false discoveries using layered critical values as proposed by Webb [7]. Layered adjustment of critical values aims to reduce the level of strictness, especially on shorter rules that are often rejected by statistical significance tests using direct adjustment of critical value. The following adjustment is used in layered adjustment:

$$\alpha_L' = \frac{\alpha}{L \times H_L}$$

where L is the current level of search, H_L is the number of all hypotheses that should be tested on level L, and α is the default critical value, usually 0.05 is used. In our case we should also take into account that each itemset is tested $a_{max} - a_{min}$ times when searching for the age interval with the highest z-value.

3. EXPERIMENTAL SETUP

The proposed age group related itemset discovery was experimentally applied to the National Hospital Discharge Survey (NHDS) data. The dataset contains hospital discharge records for approximately 1% of US hospitals. Only short-stay hospitals (hospitals with an average length of stay for all patients of less than 30 days) with at least six beds or those whose specialty is general (medical or surgical) or children's general were included in the survey. In our study we used data from 5 consecutive years from 2005 to 2009 that were joined in a single dataset containing 1,445,133 patients. Removing patients aged less than 18 years old reduced the number of records to 1,185,477. Due to specifics of gender related diagnoses, we split the dataset in two datasets based on gender of the patients. The subset with male patients contained 462,521 (39.02%) patients, while the second subset with female patients contained the remaining 722,956 (60.98%) female patients.

Each NHDS record contains the personal characteristics of the patient, including birth date or age, gender, race, and marital status; administrative information, including length of stay, and discharge status; medical information, including diagnoses, surgical and nonsurgical procedures. Each patient can have up to 7 diagnoses with an additional admitting diagnosis, where available. Age, gender and all diagnosis codes were the only attributes that were used in our study. The International Classification of Diseases, 9th Revision, Clinical Modification, or ICD-9-CM was used for coding diagnoses. ICD-9-CM coding uses taxonomy of five-digit codes, where the first three digits represent the general diagnosis and are followed by two additional digits describing a more detailed subgroup of the general diagnosis. Using very detailed five-digit codes would result in itemsets with extremely low support even with over one million patients in the dataset. Therefore all five digit codes were collapsed to more general three digit codes that still represent a wide variety of medical conditions.

3.1 Human Disease Networks

To speed up the process of searching for interesting itemsets, we propose the technique where human disease networks are used in the search process. There are two basic concepts that are usually used when constructing human disease networks: morbidity, representing the support for a single diagnosis in the given population; and co-morbidity, the support for co-occurrence of two diseases. In our experiments we compare the efficiency of three co-morbidity measures: weight (by Steinhaeuser and Chawla [6]), relative risk (by Hidalgo et al. [4]) and phi (by Hidalgo et al. [4]). Weight of edge, connecting diseases *i* and *j*, can be calculated as follows:

$$W_{ij} = \frac{C_{ij}}{M_i + M_j}$$

where C_{ij} is co-morbidity of two compared diseases and M is morbidity or prevalence of a single disease. Weight measure aims to balance the high values for more frequent co-morbidities by dividing their number by a sum of single disease prevalence.

The relative risk measure is similar to weight, but also includes the total number of patients in the population (N) and is defined as:

$$RR_{ij} = \frac{C_{ij}N}{M_iM_j}$$

Relative risk measure is intrinsically biased towards overestimation of relationships between rare diseases and underestimates the co-morbidity of more frequent diseases. This bias can be reduced by introduction of a ϕ-correlation measure, defined as:

$$\phi_{ij} = \frac{C_{ij}N - M_iM_j}{\sqrt{M_iM_j(N - M_i)(N - M_j)}}$$

All three measures were experimentally used in the interesting itemset discovery process to guide the beam search described in the following section.

3.2 Screening the Data

In our experiments a supervised beam search was used for discovery of interesting itemsets. The user sets the target diagnosis that is always used in an itemset that can be extended by any other diagnosis in the dataset. Heuristic beam search using neighbor nodes of target diagnosis in human disease network is then used to build interesting itemsets. A user defined number of neighbor nodes, with the highest co-morbidity measures, are evaluated in each iteration. Only the user defined percentage of the best items are then used for further extension in the following iteration. The itemset complexity is limited by search depth parameter. After each iteration, an itemset with the highest z-value, described in section 2, is added to a pool of final itemsets that are initially sorted by z-value. It has to be noted that beam search is a very intuitive way to search for itemsets in a network since both, the network and the itemset beam search tree are special representations of a graph.

4. RESULTS

In the first experiment, we evaluate the effectiveness of the three co-morbidity measures to lower the time complexity of the itemset discovery process. The essential hypertension (diagnosis code 401) was used as a target diagnosis due to the high prevalence of this disease in the western world. High support for this diagnosis also allows higher complexity of itemsets. However, even with relatively high support for the target diagnosis one should note that itemsets with four or more items will rapidly loose support in comparison to less complex itemsets with two or three diagnoses.

Figure 1 shows sorted z-values for top 50 itemsets derived using edge weight, relative risk and ϕ-correlation co-morbidity measures. All itemsets were calculated for age groups from 30 to 80 years with the age window set to 10 years. Number of evaluated diagnoses per iteration was set to 10 with the best 80% being used for further extension. Rule complexity was limited at 3 items that were allowed in a single itemset. Additionally a simulation of exhaustive search was performed by setting the number of evaluated items to 100 and percentage of top ranked diagnoses used to 100%. From Figure 1, we can observe that edge weight and ϕ-correlation produced rules with higher z-values, while relative risk left out a lot of potentially interesting rules. Obviously, the semi-exhaustive approach found even more rules, but the search space was almost 200 times larger compared to co-morbidity guided search described above.

It is interesting to observe the top-rated rules that were very similar in all approaches. Figure 2 represents the z-values (using ϕ-correlation) of 10 year intervals for age groups between [30, 39] and [80, 89] years for three different itemsets with the highest z-values.

It can be observed that the age group that is most exposed in itemset 401 AND 305, significantly differs from the age group of the most exposed group in target class (diagnosis 401). Translating the codes (Table 1), we can find the highest proportion of essential hypertension patients in the age group [59-68]. When essential hypertension co-occurs with nondependent abuse of drugs the age window shifts to the age group [45-54].

In Figure 2 it is possible to observe only itemsets with two items, however there are multiple additional itemsets that were discovered for essential hypertension with three items. The first such rule can be found in seventh place observing the ranked list of itemsets by their level of significance. Additionally it is also possible to compare the most exposed age groups for combinations of diagnoses between men and women. For example, the most essential hypertension prone age group for men is the interval [59-68] compared to [73-82] for women, demonstrating the huge gap between the most exposed age groups.

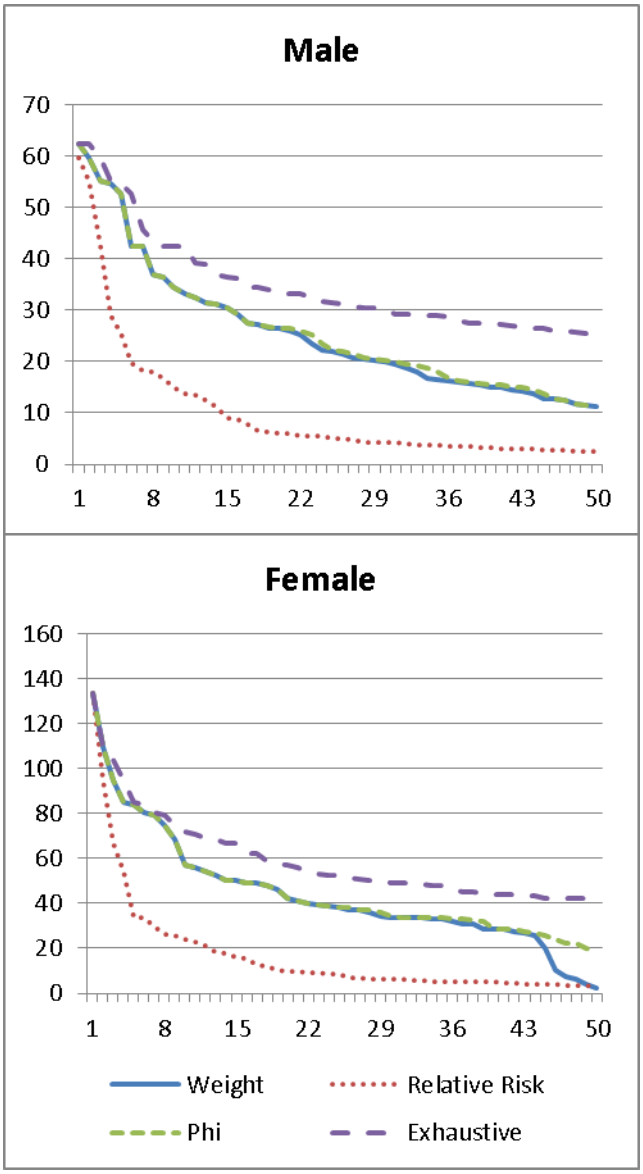

Figure 1. Z-values for 50 top ranked itemsets with diagnosis 401 (essential hypertension) as target class

On the other hand, our method uncovered an additional itemset with three items where the interval for men stays the same as for essential hypertension while the age group for women changes from [73-82] to [62-71] age group. This extreme shift occurs when hypertension co-occurs with disorders of lipoid metabolism and diabetes mellitus at the same time. Similar age group shifts

could be very important information for policymakers and insurance companies to target their campaigns or activities at different age groups.

5. CONCLUSIONS

This paper presents an itemset discovery approach for identification of age group specific patterns from hospital discharge data. We evaluate three different measures from human disease network theory to reduce the time complexity of the search process. Data collection from a five year period was used to experimentally build the interesting itemsets. Although we propose a statistical significance testing approach, we believe that the end user should decide which of the itemsets are interesting for further exploration and more rigorous statistical tests.

To further exploit the human disease networks that have to be built at the beginning of the search process, it would also be possible to use them in the final stage for visualization of obtained results. Some modern tools for visualization of dynamic networks, like Gephi [2], already allow visualization of time dependent networks.

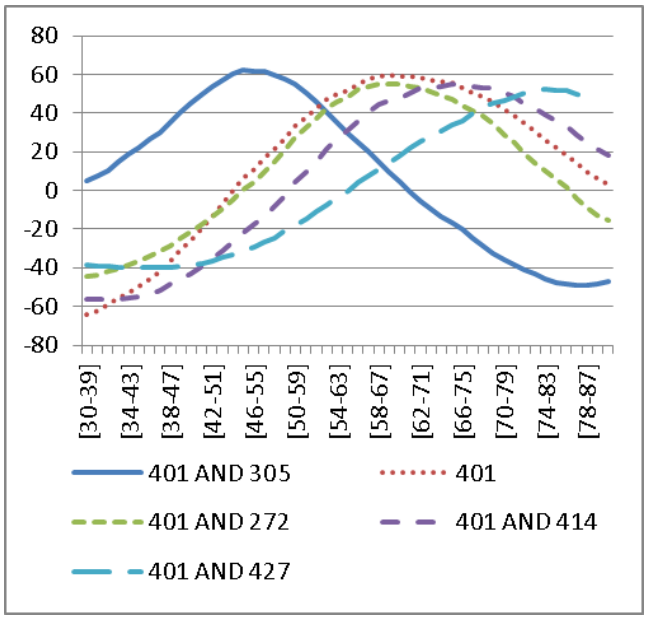

Figure 2. Results of z-test (z-value) for top five itemsets in male dataset. Curve peaks identify age groups with the highest proportion of co-morbidity for each itemset. See table 1 for explanation of diagnosis codes.

In our case the target node should be fixed and the age component used instead of time for effective visualization of disease and co-morbidity development in different age groups. Further improvements would also be possible if we would test the statistical significance of the difference in proportions between two itemsets in a group of itemsets from the same target disease group.

Table 1. Diagnosis code descriptions for diagnoses used in Figure 2.

Diagnosis code	Description
401	Essential hypertension
305	Nondependent abuse of drugs
272	Disorders of lipoid metabolism
414	Other forms of chronic ischemic heart disease
427	Cardiac dysrhythmias

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in massive databases. In Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data. Washington, DC, 207–216, 1993.

[2] M. Bastian, S. Heymann, M. Jacomy. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media, 2009.

[3] R.J. Bayardo, and R. Agrawal. Mining the most interesting rules. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99). 145–154, 1999.

[4] C.A. Hidalgo, N. Blumm, A. Barabási, N.A. Christakis. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol, 5, 2009

[5] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In Knowledge Discovery in Databases, G. Piatetsky-Shapiro and J. Frawley, Eds. AAAI/MIT Press, Menlo Park, CA., 229–248, 1991.

[6] K. Steinhaeuser and N.V. Chawla. A Network-Based Approach to Understanding and Predicting Diseases. Social Computing, Behavioral Modeling, and Prediction, Springer, 209-216, 2009.

[7] G.I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. Machine Learning, 71(2-3):307–323, 2008.

[8] G.I. Webb. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. Transactions on Knowledge Discovery from Data, 4(1), 2010.